

Differential Calculus, Tensor Products, and the Importance of Notation

Jonathan H. Manton*

August 2, 2012

Abstract

Differential calculus is used routinely across the sciences, albeit with differences in notation. These differences are especially apparent when working in higher dimensions with higher orders of derivatives. This article scrutinises an efficient coordinate-free notation, hoping to facilitate its broader adoption. Tensor products, whose purpose has been considered difficult to motivate quickly in elementary ways, are purposely shown to arise naturally in this context.

1 Introduction

The derivative of a function and the chain-rule formula for differentiating the composition of two functions are generally considered elementary because they are taught early on to students. Nevertheless, a plethora of articles exist on the chain-rule alone, including [6–9, 14, 16, 18], with [8] pointing out an error in a well-regarded book [2, p. 3]. The existence of differentiable yet nowhere monotone functions [3], while true, is far from obvious. The history is not straightforward either; Faà di Bruno was neither the first to state nor prove the higher-order chain-rule formula that bears his name [5, 10].

The present article elucidates that even constructing a convenient notation is not entirely elementary. It propounds a minor yet simplifying modification of the “ Df ” notation for Fréchet derivatives [11, Chapter 8] in certain situations.

The Df notation is not prevalent in applied fields that favour instead gradients, Jacobians and Hessians [13]. This is somewhat surprising since Section 2 exemplifies the convenience of the Df notation over element-wise differentiation.

Subtleties arise when differentiating abstract expressions such as $D^3(f \circ g)$. Section 4 highlights the safe approach is tedious while the common alternative of omitting variables requires care. The cause of notational difficulty is that higher-order derivatives evaluated at a point are multi-linear maps whose arguments can again be multi-linear maps, resulting in a tree structure with functions within functions. The tensor product suggests itself as a way of collapsing the tree to a linear structure by converting multi-linear maps to linear maps.

The details of how to use tensor products to simplify working with Fréchet derivatives are not readily found in the literature. No mention is made in

*Department of Electrical and Electronic Engineering, The University of Melbourne, Victoria 3010 Australia. Email: j.manton@ieee.org

the following textbooks on differential calculus [1, Chapter 2], [11, Chapter 8], [20, Chapter 4], [21, Chapter 5], nor in the following textbooks on differential geometry [2, Chapter 1], [4, Chapter I.2], [12, Chapter I.3], [15].

The tensor product can be difficult to motivate early in an undergraduate curriculum due to a shortage of elementary contexts genuinely requiring a tensor product. Introducing it as a part of calculus changes this; students initially treat \otimes as a formal symbol separating the arguments of a function and become familiar with its product-like behaviour, then later appreciate it reduces multi-linear maps to linear maps.

2 An Example in Matrix Space

The Df notation provides a coordinate-free approach to differential calculus in matrix spaces. It is presented here by way of example.

Consider $f(X) = \text{tr}\{X^T A X\}$ where $\text{tr}\{\}$ denotes trace, superscript T denotes transpose, and A and X are matrices of compatible dimensions. Often the derivative of such a function f is represented by its Jacobian matrix whose ij -th element is the partial derivative of f with respect to the element X_{ij} of X . Evaluating these partial derivatives from first principles is straightforward but tedious: use $(AB)_{ij} = \sum_k A_{ik} B_{kj}$ twice and $\text{tr}\{Z\} = \sum_i Z_{ii}$ to obtain $f(X) = \sum_{ijk} X_{ji} A_{jk} X_{ki}$, differentiate normally, and attempt to convert the answer back to matrix form.

The following is an alternative approach. Explanations follow in subsequent sections. Fix a matrix Z of the same dimensions as X . Then:

$$f(X + tZ) - f(X) = \text{tr}\{(X + tZ)^T A (X + tZ)\} - \text{tr}\{X^T A X\} \quad (1)$$

$$= (\text{tr}\{Z^T A X\} + \text{tr}\{X^T A Z\})t + (\text{tr}\{Z^T A Z\})t^2. \quad (2)$$

Derivatives represent linear approximations, and (2) shows the derivative of f at X in the direction Z is $\text{tr}\{Z^T A X\} + \text{tr}\{X^T A Z\}$. The meaning may not be clear yet, but the calculation was simple!

The mapping $Z \mapsto \text{tr}\{Z^T A X\} + \text{tr}\{X^T A Z\}$ is linear: if it sends Z_1 to c_1 and Z_2 to c_2 then it sends $\alpha Z_1 + \beta Z_2$ to $\alpha c_1 + \beta c_2$ for $\alpha, \beta \in \mathbb{R}$. This linear mapping is the (Fréchet) derivative of f .

$$Df(X) \cdot Z = \text{tr}\{Z^T A X\} + \text{tr}\{X^T A Z\} \quad (3)$$

$$= \text{tr}\{Z^T (A + A^T) X\}. \quad (4)$$

The Jacobian matrix can be read off as $(A + A^T)X$.

Treating Z as a constant and differentiating (4) gives

$$(D^2 f(X) \cdot Z) \cdot T = \text{tr}\{Z^T (A + A^T) T\}. \quad (5)$$

The Hessian is $(A + A^T)$. The left-hand side of (5) is more commonly written as $D^2 f(X) \cdot (Z, T)$.

3 First-Order Derivatives and Gradients

The definition $f'(x) = \lim_{t \rightarrow 0} t^{-1}[f(x+t) - f(x)]$ of the derivative of a function $f : \mathbb{R} \rightarrow \mathbb{R}$ extends in several ways to functions $f : U \rightarrow V$ between finite-

dimensional vector spaces U and V . The reader may take, for concreteness, U and V to be scalars \mathbb{R} , vectors \mathbb{R}^n or matrices $\mathbb{R}^{n \times m}$.

One extension considers directional derivatives so as to reduce to the case $g : \mathbb{R} \rightarrow V$, $g(t) = f(x + tz)$ for fixed $x, z \in U$, for which the same formula as above can be used:

$$D_z f(x) = \lim_{t \rightarrow 0} \frac{f(x + tz) - f(x)}{t}. \quad (6)$$

If the limit exists for all z then (6) is called the Gâteaux derivative of f at x .

Another extension looks beyond (6) and focuses on the geometric meaning of $f'(x)$ as the gradient of the line of best fit to the graph of f at x . This suggests defining the derivative as the best linear approximation of f at x . Precisely, fix x and assume there exists a linear function $A_x(z)$ such that

$$\lim_{z \rightarrow 0} \frac{\|f(x + z) - f(x) - A_x(z)\|}{\|z\|} = 0. \quad (7)$$

Then A_x is unique and is called the Fréchet derivative of f at x , denoted $Df(x)$. Sometimes, evaluation in a particular direction is denoted using a dot, as in (3). That is, $Df(x) \cdot z = A_x(z)$.

The limit in (7) must exist for *any* sequence $\{z_n\}_{n=1}^\infty$ with $z_n \rightarrow 0$. Even if the mapping $z \mapsto D_z f(x)$ in (6) is linear for a fixed x , the Fréchet derivative need not exist because it is possible for (7) to hold for sequences z_n converging to the origin along straight lines but not for sequences following certain curved trajectories. This occurs when the limit is not uniform across straight lines: convergence to zero is fast along some lines but arbitrarily slow along others.

An expedient technique for calculating Fréchet derivatives is guess-then-verify. Verification is unnecessary if the Fréchet derivative is known to exist by other means. In Section 2, f is a polynomial, hence the Fréchet derivative exists and can be found using directional derivatives, either explicitly as in (2) or, in more complicated situations, by using truncated Taylor series approximations. Of course, tables and rules could be used instead.

If $f : U \rightarrow \mathbb{R}$ is a scalar function then its gradient at x is defined with respect to an inner product. This is often forgotten because the Euclidean inner product is chosen without mention in many textbooks. In matrix space, the Euclidean inner product is $\langle A, B \rangle = \text{tr} \{B^T A\}$. For a fixed matrix G , $A(Z) = \langle G, Z \rangle$ is a linear functional, and every linear functional can be written this way. The gradient of f at X is the matrix G_X such that $Df(X) \cdot Z = \langle G_X, Z \rangle$.

4 Second-order Derivatives and Hessians

The Fréchet derivative of $f : U \rightarrow V$ is $Df : U \rightarrow L(U; V)$ where $L(U; V)$ is the vector space of all linear maps from U to V . Applying D to Df yields the second-order derivative $D^2 f : U \rightarrow L(U; L(U; V))$. A second-order derivative requires not one, but two, directions: $(D^2 f(X) \cdot T) \cdot Z$. The right-hand side of (11) interprets this as the rate of change in the direction T of the directional derivative $Df(X) \cdot Z$.

To the letter of the law, $D^2 f(X)$ is calculated from (4) as follows. Working directly with $Df(X) \cdot Z$ is not allowed because $Df(X)$ must be treated as an element of $L(U; V)$ when computing $D^2 f(X) \cdot T = D(Df)(X) \cdot T$. By assuming

the Fréchet derivative exists, it suffices to work with directional derivatives:

$$D(Df)(X) \cdot T = \lim_{t \rightarrow 0} \frac{Df(X + tT) - Df(X)}{t}. \quad (8)$$

For clarity, let $L_t = Df(X + tT) \in L(U; V)$. For fixed t , both $L_t - L_0$ and $(L_t - L_0)t^{-1}$ are linear operators in $L(U; V)$. The vector space structure on $L(U; V)$ is that induced by pointwise operations: $(L_t - L_0)t^{-1}$ evaluated at Z is $(L_t \cdot Z - L_0 \cdot Z)t^{-1}$ by definition. A sequence of linear operators converges if and only if it converges pointwise (throughout, all vector spaces are finite-dimensional). Thus, the right-hand side of (8) can be determined pointwise:

$$\left(\lim_{t \rightarrow 0} \frac{Df(X + tT) - Df(X)}{t} \right) \cdot Z = \lim_{t \rightarrow 0} \frac{Df(X + tT) \cdot Z - Df(X) \cdot Z}{t} \quad (9)$$

$$= \text{tr} \{ Z^T (A + A^T) T \}. \quad (10)$$

In words, $D(Df)(X) \cdot T$ is the linear operator $Z \mapsto \text{tr} \{ Z^T (A + A^T) T \}$.

A nominally different quantity is the derivative $Dg(X) \cdot T$ where $g(X) = Df(X) \cdot Z$ for a fixed Z . Nevertheless, $Dg(X) \cdot T = \text{tr} \{ Z^T (A + A^T) T \}$, the same as (10). Indeed, the pointwise vector space structure on $L(U; V)$ means

$$(D^2f(X) \cdot T) \cdot Z = (D(Df)(X) \cdot T) \cdot Z = D(Df \cdot Z)(X) \cdot T. \quad (11)$$

Therefore D^2f can be computed from $Df(X) \cdot Z$ by treating Z as a constant and differentiating with respect to X . This is how (5) is obtained from (4).

The above notation is simple but cumbersome. Textbooks generally drop the variables, writing the chain rule and product rule as

$$D(f \circ g) = (Df \circ g) Dg, \quad (12)$$

$$D(fg) = (Df)g + fDg. \quad (13)$$

Direct application can lead to confusion though:

$$D^2(f \circ g) = D((Df \circ g) Dg) \quad (14)$$

$$= (D(Df \circ g)) Dg + (Df \circ g) D^2g \quad (15)$$

$$= ((D^2f \circ g) Dg) Dg + (Df \circ g) D^2g. \quad (16)$$

Taken literally, it is a nonsense to multiply $D^2f \circ g$ with Dg twice. Only with experience can $(D^2(f \circ g)(X) \cdot T) \cdot Z$ be deduced from (16).

Including directions from the start reveals

$$D(f \circ g) \cdot Z = (Df \circ g) \cdot (Dg \cdot Z), \quad (17)$$

$$D(f \cdot (g \cdot Z)) \cdot T = (Df \cdot T) \cdot (g \cdot Z) + f \cdot ((Dg \cdot T) \cdot Z), \quad (18)$$

$$\begin{aligned} (D^2(f \circ g) \cdot T) \cdot Z &= ((D^2f \circ g) \cdot (Dg \cdot Z)) \cdot (Dg \cdot Z) \\ &\quad + (Df \circ g) \cdot ((D^2g \cdot T) \cdot Z). \end{aligned} \quad (19)$$

Here, X is omitted because it is simple enough to feed it in to the terms requiring it. To be clear, $Df \circ g$ means evaluate Df at $g(X)$.

Neither approach is particularly friendly. The former omits important details while the latter is tedious; the reader is invited to derive (19) from either (17) and (18), or from (12) and

$$D(fg) \cdot Z = (Df \cdot Z)g + f(Dg \cdot Z). \quad (20)$$

For scalar fields $f : U \rightarrow \mathbb{R}$, the unique linear operator H_X satisfying $(D^2f(X) \cdot T) \cdot Z = \langle H_X \cdot T, Z \rangle$ is the Hessian of f at X . The ordering is unimportant because $D^2f(X)$ is symmetric: $(D^2f(X) \cdot T) \cdot Z = (D^2f(X) \cdot Z) \cdot T$ for all Z and T . When the Euclidean inner product is used, H_X agrees with what is called the Hessian matrix [13].

5 A Tensor Product Notation for Derivatives

Given $f : U \rightarrow L(V; W)$ and $g : U \rightarrow L(Y; V)$, Df maps into $L(U; L(V; W))$ whereas g maps into $L(Y; V)$, indicating technically the product $(Df)g$ cannot be formed. The tensor product allows replacing $L(U; L(V; W))$ by $L(U \otimes V; W)$. Equation (13) becomes

$$D(fg) = (Df)(I \otimes g) + fDg \quad (21)$$

where I is the identity map. This has the correctness of (20) and almost the same brevity as (13).

The obvious role of the tensor product is directing variables to their correct targets: the g in $(Df)g$ blocks Z from reaching Df when applied on the right, while the I in $Df(I \otimes g)$ allows the Z through. Although the direct sum would serve equally well in this role, it is the tensor product that behaves correctly under differentiation:

$$D(f \otimes g) = (Df \otimes g) + (f \otimes Dg). \quad (22)$$

In particular, (21) can be differentiated again by using $D(I \otimes g) = (DI \otimes g) + (I \otimes Dg) = (0 \otimes g) + (I \otimes Dg) = I \otimes Dg$.

If f is itself a derivative then (12) becomes

$$D(f \circ g) = (Df \circ g)(Dg \otimes I). \quad (23)$$

Following these rules gives

$$D^2(f \circ g) = (D^2f \circ g)(Dg \otimes I)(I \otimes Dg) + (Df \circ g)D^2g \quad (24)$$

$$= (D^2f \circ g)(Dg \otimes Dg) + (Df \circ g)D^2g, \quad (25)$$

$$D^3(f \circ g) = (D^3f \circ g)(Dg \otimes Dg \otimes Dg) + (D^2f \circ g)[(D^2g \otimes Dg) + 2(Dg \otimes D^2g)] + (Df \circ g)D^3g. \quad (26)$$

The remainder of this section gives the intermediate steps. Section 6 presents a formal description of the notation.

Start with $D(f \circ g) = (Df \circ g)Dg$. Differentiate to get $D(Df \circ g)(I \otimes Dg) + (Df \circ g)D^2g$. This time, (23) is required: $D(Df \circ g) = (D^2f \circ g)(Dg \otimes I)$. Tensor products of linear maps satisfy the rule $(A \otimes B)(C \otimes D) = (AC \otimes BD)$. Therefore, $(Dg \otimes I)(I \otimes Dg) = Dg \otimes Dg$.

To obtain (26), first apply the product rule (21) to the two additive terms in (25). Note $D(Dg \otimes Dg) = (D^2g \otimes Dg) + (Dg \otimes D^2g)$. At this point,

$$\begin{aligned} D^3(f \circ g) &= (D^3f \circ g)(Dg \otimes I)(I \otimes (Dg \otimes Dg)) \\ &\quad + (D^2f \circ g)[(D^2g \otimes Dg) + (Dg \otimes D^2g)] \\ &\quad + (D^2f \circ g)(Dg \otimes I)(I \otimes D^2g) + (Df \circ g)D^3g. \end{aligned} \quad (27)$$

The first I in (27) acts on $U \otimes U$ whereas the second acts on U . Regardless, it is agreeable to equate $(Dg \otimes I)(I \otimes (Dg \otimes Dg))$ with $Dg \otimes (Dg \otimes Dg) = Dg \otimes Dg \otimes Dg$, and (26) readily follows from (27).

6 Formal Description

The tensor product notation used in Section 5 is stated formally below. Some intricacies appear, but go unnoticed in practice. The tensor product is generally not required when differentiating a given function; cf., Section 2. It can simplify the differentiation by hand of abstract expressions, such as when seeking bounds like the one in (38).

All spaces are finite-dimensional vector spaces. Basic properties of tensor products are used [19]. The main principle is that canonical isomorphisms of vector spaces can be applied freely because they essentially commute with the Fréchet derivative.

Given $f : U \rightarrow V$, define $\bar{D}^k f : U \rightarrow L(U \otimes \cdots \otimes U; V)$ by

$$\bar{D}^k f(X) \cdot (Z_1 \otimes \cdots \otimes Z_k) = ((D^k f(X) \cdot Z_1) \cdots) \cdot Z_k. \quad (28)$$

For $g : U \rightarrow L(V; W)$, define $\bar{D}_V^k(g) : U \rightarrow L(U \otimes \cdots \otimes U \otimes V; W)$ by

$$\bar{D}_V^k(g) \cdot (Z_1 \otimes \cdots \otimes Z_k \otimes T) = (((D^k g(X) \cdot Z_1) \cdots) \cdot Z_k) \cdot T. \quad (29)$$

Although $\bar{D}^2 f \neq \bar{D}(\bar{D}f)$, they agree up to a canonical linear isomorphism. In fact, $\bar{D}^k f = \bar{D}_U^{k-1}(\bar{D}f)$. For all intents and purposes, $\bar{D}_V^2(g)$ agrees with $\bar{D}_{U \otimes V}(\bar{D}_V(g))$ because applying the canonical identification $U \otimes (U \otimes V) \cong U \otimes U \otimes V$ in practice simply means omitting a pair of brackets.

Given $g : U \rightarrow L(V; W)$ and $h : U \rightarrow L(W; Y)$, the product rule is

$$\bar{D}_V(hg) = \bar{D}_W(h) (I_U \otimes g) + h \bar{D}_V(g) \quad (30)$$

where $I_U : U \rightarrow U$ is the identity map and $I_U \otimes g$ is a tensor field over U whose value at $X \in U$ is $I_U \otimes (g(X)) \in L(U \otimes V; U \otimes W)$.

Related is the application of a linear map to a vector; given $f : U \rightarrow W$ then

$$\bar{D}(h \cdot f) = \bar{D}_W(h) (I_U \boxtimes f) + h \bar{D}f \quad (31)$$

where $(I_U \boxtimes f)(X)$ is the linear map $Z \mapsto (Z \otimes f(X))$. Later, by minor abuse of notation, \otimes will replace \boxtimes . Since $L(U; V) \boxtimes W = L(U; V \otimes W) \cong L(U; V) \otimes W$, both \boxtimes and \otimes behave essentially the same way when differentiated.

For $d : V \rightarrow L(W; Y)$, $e : U \rightarrow V$ and $f : V \rightarrow W$, the two chain rules are

$$\bar{D}(f \circ e) = (\bar{D}f \circ e) \bar{D}e, \quad (32)$$

$$\bar{D}_W(d \circ e) = (\bar{D}_W(d) \circ e) (\bar{D}e \otimes I_W) \quad (33)$$

where $I_W : W \rightarrow W$ is the identity map.

For $e : U \rightarrow V$ and $f : U \rightarrow W$, the tensor product rule is

$$\bar{D}(e \otimes f) = (\bar{D}e \boxtimes f) + (e \boxtimes \bar{D}f) \quad (34)$$

where \boxtimes combines a vector and a linear map to form a linear map, as in (31). For $g : U \rightarrow L(V; W)$ and $h : U \rightarrow L(C; Y)$, the tensor product rule is

$$\bar{D}_{V \otimes C}(g \otimes h) = (\bar{D}_V(g) \otimes h) + (g \otimes \bar{D}_C(h)). \quad (35)$$

For $e : U \rightarrow V$ and $h : U \rightarrow L(C; Y)$,

$$\bar{D}_C(e \boxtimes h) = (\bar{D}e \otimes h) + (e \boxtimes \bar{D}_C(h)). \quad (36)$$

If the codomains of all functions are spaces of linear maps then the situation is particularly simple; (30), (33) and (35) suffice. This is the typical situation when computing higher-order derivatives because the codomain of the derivative of a function is a space of linear maps. It is possible to reduce to this situation by replacing $f : U \rightarrow W$ with $\tilde{f} : U \rightarrow L(\mathbb{R}; W)$ where $f(X) = \tilde{f}(X) \cdot 1$. The $\cdot 1$ can be removed, the derivatives calculated, and the $\cdot 1$ applied at the very end. This explains the similarity of (30) and (31), and of (34), (35) and (36).

In practice, it is easier to replace \boxtimes by \otimes than replace f by \tilde{f} . No confusion arises because \boxtimes and \otimes behave the same way with respect to addition, multiplication and differentiation.

The subscripts on \bar{D} used in (30)–(36) merely keep all derivatives in a consistent form and can be dropped. When computing higher-order derivatives recursively, to account for \bar{D}^k differing from \bar{D}^{k-1} by a linear isomorphism, it is only necessary to remove any remaining brackets in tensor products at the end of each step, e.g., replace $Dg \otimes (Dg \otimes Dg)$ by $Dg \otimes Dg \otimes Dg$ in (27).

Once \boxtimes is replaced by \otimes and the subscripts dropped on \bar{D} , the rules collapse to those in Section 5.

7 Discussion

Section 6 elicited the relationship between \bar{D} and D . This viewpoint highlights the usefulness of the tensor product as a mathematical operation and validates its use when computing bounds on the (operator) norms of derivatives, e.g.,

$$\|D^2(f \circ g)\| \leq \|D^2 f \circ g\| \|Dg \otimes Dg\| + \|Df \circ g\| \|D^2 g\| \quad (37)$$

$$\leq \|D^2 f \circ g\| \|Dg\|^2 + \|Df \circ g\| \|D^2 g\|. \quad (38)$$

Alternatively, a mechanical calculus can be developed, where \otimes is a formal symbol used to direct variables to their correct targets. This would follow the course of building a class Ω of allowable expressions, explaining how D is applied to members of this class, and verifying the class is algorithmically closed under D . This mechanical viewpoint could be used to write trees of nonlinear functions of multiple arguments as a serial composition of functions of a single argument; the utility is questionable though.

Fréchet derivatives are defined on Banach spaces. The aforementioned mechanical viewpoint implies the tensor product notation remains applicable. Depending on the application, a subtlety is that tensor products are not uniquely defined on Banach spaces; different choices of norms, and hence completions with respect to that norm, are possible [17].

8 Conclusion

The presentation aimed to complement traditional texts on matrix differential calculus. The Df notation is convenient for differentiating given functions (Section 2) but has its subtleties when differentiating abstract expressions (Section 4). Tensor products provide a notational convenience that simplifies certain

calculations (Section 5). This is pedagogically interesting as an elementary yet genuine application of the tensor product.

References

- [1] R. Abraham, J. E. Marsden, and T. Ratiu. *Manifolds, Tensor Analysis, and Applications*, volume 75 of *Applied Mathematical Sciences*. Springer, second edition, 1988.
- [2] R. Abraham and J. Robbin. *Transversal Mappings and Flows*. W. A. Benjamin, Inc, 1967.
- [3] A. M. Bruckner, J. Mařík, and C. E. Weil. Some aspects of products of derivatives. *The American Mathematical Monthly*, 99(2):134–145, 1992.
- [4] W. L. Burke. *Applied Differential Geometry*. Cambridge University Press, 1985.
- [5] A. D. D. Craik. Prehistory of Faà di Bruno’s formula. *Amer. Math. Monthly*, 112(2):119–130, 2005.
- [6] A. Dresden. The derivatives of composite functions. *The American Mathematical Monthly*, 50(1):9–12, 1943.
- [7] H. Flanders. From Ford to Faà. *The American Mathematical Monthly*, 108(6):559–561, 2001.
- [8] L. E. Fraenkel. Formulae for high derivatives of composite functions. *Math. Proc. Cambridge Philos. Soc.*, 83(2):159–165, 1978.
- [9] T. J. Grilliot. Classroom Notes: Derivatives of Composite Functions. *Amer. Math. Monthly*, 69(9):912–914, 1962.
- [10] W. P. Johnson. The curious history of Faà di Bruno’s formula. *The American Mathematical Monthly*, 109(3):217–234, 2002.
- [11] J. Jost. *Postmodern Analysis*. Universitext. Springer, second edition, 2003.
- [12] S. Lang. *Fundamentals of Differential Geometry*. Number 191 in Graduate Texts in Mathematics. Springer, 1999.
- [13] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley Series in Probability and Statistics. John Wiley and Sons, revised edition, 1999.
- [14] M. McKiernan. On the n th derivative of composite functions. *Amer. Math. Monthly*, 63:331–333, 1956.
- [15] J. R. Munkres. *Analysis on manifolds*. Addison-Wesley Publishing Company Advanced Book Program, Redwood City, CA, 1991.
- [16] J. Riordan. Derivatives of composite functions. *Bull. Amer. Math. Soc.*, 52:664–667, 1946.

- [17] R. A. Ryan. *Introduction to tensor products of Banach spaces*. Springer Monographs in Mathematics. Springer-Verlag London Ltd., London, 2002.
- [18] K. P. Rybakowski. Formulas for higher-order Fréchet derivatives of composite maps, implicitly defined maps and solutions of differential equations. *Nonlinear Analysis, Theory, Methods and Applications*, 16(6):517–532, 1991.
- [19] T. Yokonuma. *Tensor Spaces and Exterior Algebra*. Number 108 in Translations of Mathematical Monographs. American Mathematical Society, 1992.
- [20] E. Zeidler. *Nonlinear functional analysis and its applications. I*. Springer-Verlag, New York, 1986.
- [21] V. A. Zorich. *Mathematical Analysis I*. Universitext. Springer, 2004.